THE REPRESENTATION IN MODELS OF WHAT SPEAKERS KNOW

Celia Scully

University of Leeds
Leeds LS2 9JT, U.K.

ABSTRACT

The need to include in models of speech production all the physical processes and the interactions between them is discussed. The role of trial and error with auditory monitoring in learning to achieve the goals of speech production is emphasised, for models as well as for real speakers.

1. INTRODUCTION

Perkell's paper focusses on global models which map from discrete linguistic units through to a synthetic acoustic signal. The need to take the mapping further still, to speech perception and lexical access, is emphasised. This is an extremely important aspect of production modelling. Separation of control and "plant" is part of the philosophy of some global models. It may be interesting to consider where the line should be drawn between these two components. Finally, I should like to consider the nature of the problems of including biomechanical systems in a composite model and consider some priorities in the tasks confronting us.

2. MAPPING ALL THE WAY THROUGH TO PERCEPTION

Real speech production is goal-directed sensori-motor behaviour. The goal is a sequence of auditory patterns, broadly enough defined to accommodate variations such a cross-speaker differences; there are many different combinations of acoustic patterns that are acceptable as a specified goal. The goal is a different one, depending upon the context and style of the speech. Speakers learn to make auditory patterns during speech acquisition, and skilled (normal adult) speakers have already learned and stored the appropriate neural patterns for force and time for achieving each auditory goal.

The actions are the means to the end, not the goal itself. Extensive learning can be expected to give the speaker, like the musician or the games player, a huge repertoire of patterns of activity. The linguistic structures posited as inputs to a global model of speech production can be considered as a (partial) description of the auditory goal itself. Acquisition of speech must surely include, and indeed rely heavily upon, trial and error with monitoring, by auditory combined with other kinds of sensory feedback.

It must be acknowledged that the neural processes which permit a speaker the flexibility to speak in many different styles are mysterious. Perhaps it is premature to try to

address that question at the present time? After all, the organisation of motor activity in the Central Nervous System (CNS) is more difficult to study than perception; and speech production must surely be one of the most complex of all motor skills.

Data and modelling for the transform from acoustics to perception constitute an enormously important research area in its own right. I do not believe that it is central to the immediate tasks confronting us in speech production research, however. The perfect mapping device exists close at hand - that is, ordinary speakers of the language concerned. They can tell the modeller whether the auditory goals have been achieved or not. The fact that we do not understand this mapping need not worry us unduly. It is not even essential to acoustically segment the synthetic speech signal or to relate particular portions of the soundwave to individual linguistic elements as part of this assessment procedure. There is, in any case, no real justification for labelling a particular acoustic event, such as the onset of voicing, as a boundary between two linguistic units such as a consonant and a vowel.

Formal perception tests on every speech or speech-like sequence generated by a model or analysed from real speech would be an excessively demanding counsel of perfection. But it is perfectly feasible for one or more native speaker-listeners to check all recordings auditorily, and to confirm that the output has indeed achieved the auditory goal set.

The role of auditory feedback in trial and error improvement of performance of real speech cannot be over-stated. Speakers with a significant hearing loss cannot be expected to learn the appropriate actions. The matter for comment here is not that they get so many aspects of the speech wrong, but that any at all are right. It is no coincidence that the early attempts at speech with a composite model of speech production can be very abnormal and indeed resemble deaf speech. So a global or other model which generates sounds should not be judged a failure after its first few attempts. It must learn through experience (and theory too, where that is available) the right combinations of gestures needed. It is a vehicle for the actions which achieve the goals set, not a model of the goals themselves.

3. SEPARATION OF CONTROL AND PLANT

The name "model" is given to two rather different kinds of endeavour. A model may be a simulation of some physical processes or it may mean a hypothesised form of organisation and control, as are the conflicting comb and chain models which state how actions relate to phoneme-type linguistic units. It seems clear that both these last two are over-simplified views of the CNS. But, even assuming more plausible models of control by the CNS, I am not sure that we are able to say at present that some aspects of speech production are the control, while other aspects are the controlled system, the plant.

The signals which activate muscles originate in the CNS where the inputs to the speech producing system also reside, but perhaps the former of these neural processes could be considered to be one of the stages of speech production which together constitute the plant, simulated by the composite model. Conceptually, the following processes may be separated and listed in order: neural, muscular, articulatory, aerodynamic, acoustic sources, acoustic filtering, and radiation of a soundwave. But we know that these separate stages interact with each other. From the proposition that skilled speakers have already learned the neural patterns needed to achieve a specific auditory goal, it follows that speakers have knowledge about all these stages and their interactions, as applied to their own speech production at least.

Many speakers are not consciously aware that their oral pressure rises a little, but only to one or two cm H_2O above atmospheric pressure, during the production of nasal and approximant consonants. But the CNS of almost any speaker knows that certain force-time combinations resulting in particular coordinations for the vocal tract closure and the velum lowering and raising are unsuccessful for the production of nasal consonants. One reason for this is that oral air pressure rises too high, and an intrusive plosive is perceived. The auditory goal specifies the right sound patterns in the right order, with no extra sounds and none omitted. So aerodynamic effects are not simply the consequences of linguistically-determined neural commands - they influence the form of the neural patterns in the first place.

This example is intended to illustrate the proposition that the neuromuscular, mechanical, aerodynamic and acoustic stages all combine to determine the form of the neural signals which cause the actions. The properties, mechanisms and constraints of all these stages are crucially important both in setting limits for the force-time plan and in generating the rich, complex details of structure and multiple acoustic cues for perception in the speech signal.

4. TIMING AND ARTICULATORY EVENTS

Much research is directed towards the timing patterns of speech. I agree very much with the view that durations of acoustic segments are not simply imposed on linguistic structures. It has been argued for many years, and the principle has been made explicit in models, that one of the strongest constraints in speech production is the time required for an individual solid structure such as the tongue or the vocal folds to be accelerated and then decelerated and so perform an articulatory transition, where that gesture is essential for the achievement of the auditory goal. This seems to act as a very important factor, interacting with the control of auditory length and prominence for vowels and for some consonants in determining speech segment durations (see, for example,[3], [8], [9], [10]). As Perkell points out, timing of actions, or of the application of forces to achieve movement, must still be specified.

Inter-articulator coordination is of the essence in speech production. Elsewhere I have suggested that some time intervals between articulatory events might perhaps be preserved across a change of speaking rate, while some actions might be dispensed with [9]. The Haskins modelling is associated with the suggestion that inter-articulator timing may be expressed as a constant phase within postulated cycles of movement. There does not seem to be any obvious basic principle, comparable to a criterion for speech production such as minimisation of work done, to favour one view or the other. Here, a model which attempts to sketch all the physical processes which constitute the mapping, either from muscle forces and their timing or from articulation, to the acoustic signal, can be neutral. It can, as Perkell says, serve as a means of focussing experiments.

5. THE BIOMECHANICS OF SPEECH PRODUCTION: COMPONENTS OF MODELS

We cannot wait for the time when the properties of all the systems are thoroughly understood; we need to incorporate them now in composite or global models of speech production. This means that we need to try to capture, qualitatively at least, some of the observed behaviour. This is, of course, a very hard task, but it must be undertaken. It may be better tactics not to focus too much effort on the input and control mechanisms, but instead take on the more modest but still very ambitious task of trying to describe, and, where possible, explain the behaviour of natural speech in all its aspects, from neuromuscular processes right through to the radiation of soundwaves. Like scientists in other disciplines, we should be content to advance our understanding little by little, piecemeal.

One severe limitation on progress is the lack of widespread availability of the advanced techniques which have been developed now for studying natural speech production. In view of the difficulties of the task, the small numbers of researchers and the limited funds available, perhaps it would be in our interests to pool our skills and resources by organising ourselves into quite large collaborative groups, with travelling speakers.

An individual researcher or group should pose the questions which interest them; we cannot and need not all aim at a global model. But it will always be important to keep sight of the implications for subsequent processes, and, especially, for the output acoustic and auditory patterns.

There is a need for a phenomenological approach to modelling at present, but better true physical models need to be developed also. One source of frustration in modelling is that the basic mathematical and physical theory for many of the processes have been so little developed.Take, for example, the conditions controlling the presence or absence and the spectral properties of turbulence noise sources. In a well established science recognised as having practical importance, surely the pioneering work of Stevens [12] would have been followed up by armies of researchers? Work on the problems of turbulence noise in jets has, I am sure, received plenty of attention in the intervening years. There is more cause for optimism now, as regards this particular example [11],[2].

6. THE INDIVIDUAL PROCESSES: INTERACTIONS AND SOME QUESTIONS

6.1. Neural Signals and Muscle Length Changes

The neural signals to the muscles interact with the muscle length changes. The timing patterns chosen can exploit these interactions to maximise force output, by ensuring that each muscle of a reciprocally acting pair is stretched prior to its innervation, a principle, which seems to be applied by fish to their swimming muscles [1] and in speech also [4]. A model of the innervation of a reciprocally acting pair of muscles, as for fast head movements [7] could perhaps generate matches to and suggest explanations for the coordination of electromyographic traces seen in speech.

6.2. Articulation

Assuming that the notion of an articulatory event is useful, the choice of particular moments as candidates is by no means self evident. Gestures, especially the middle, high velocity portions of transitions, may well be more important than the end point reached, as the thing controlled.

The positive correlation between maximum velocity near the middle of the transition and the distance transversed seems to be a strong constraint. In other non-speech tasks, subjects could not easily be made to bypass this relationship [6]. Some speech data show duration, as well as peak velocity, correlated with distance tranversed (Keller, [5] 343-364). These findings suggest the possibility that transitions large and small, fast and slow, may have kinematic similarity, with the dimensionless number $\mathbf{v^2/al}$ kept constant, where **v**, **a**, and **l** are characteristic velocities, accelerations and lengths respectively. Distance, velocity and duration might perhaps be considered as dependent variables, with muscle force as the controlling variable.

6.3. Aerodynamics

It is not really so difficult to include aerodynamic processes in models of speech production, although present representations of the processes are highly simplified. The use of packages such as the NAG (Numerical Algorithms Group) routines for numerical solution of simultaneous differential equations is to be recommended. Should the respiratory control be considered as a nett expiratory force (Ohala, [5] 23-53)? In cardiovascular studies the question has been posed as to whether the heart is a flow source or a pressure source (Pedley, personal communication). This question needs to be considered further for respiratory control in speech.

Some quite basic parameters, such as the volume enclosed in the vocal tract cavity, need to be measured, or at least estimated. Magnetic resonance and ultrasonic imaging offer hope for this as well as for the difficult task of improving the mapping from mid-sagittal views of the vocal tract onto area functions.

6.4. Acoustic sources

Individual articulatory events do not lead directly to sound patterns. It is important to consider the total effect of the movements of *all* the structures involved, including larynx and subglottal respiratory actions as well as those which shape the supraglottal vocal tract. Furthermore, articulatory geometry and aerodynamics interact to generate acoustic sources. Thus, for example, the moment of the onset of voicing for a vowel following a consonant is not a direct reflection of any one articulatory event; it depends on the interaction of at least three factors: the air pressure drop across the glottis, the state of adduction of the vocal folds, and the stiffness and effective mass of the vocal folds. So the preceding consonant will influence the moment of voice onset, for example if it requires the vocal folds to be abducted.

The need to specify articulatory parameter values at the moments of closure and release for consonants (Scully, [5] 151-186) should not be seen as a major problem. This is precisely the interest attached to the modelling. If the model generates acoustic sources in a way which approximates, however roughly, the processes of real speech, it can help to demonstrate how and why particular combinations of actions and coordination are chosen by speakers. By perturbing the timing and other aspects of the simulated articulation, modelling can investigate the limits within which the auditory goal is attained, and the covariations in multiple acoustic pattern features associated with the variability found in natural speech.

7. CONCLUSIONS

It is true that there are severe difficulties in obtaining real speech data, especially for the larynx, and that too much use has to be made of analysis-by-synthesis at present; but that is not a reason for avoiding the problems of modelling all the biomechanics of speech production. Apart from improved theory and more data, there is a real need to ease the burden for the modeller, for example by the development of code books and the use of graphical, file handling and mathematical techniques, so as to reduce the long auditory feedback loop, which puts the experimenter into almost the same situation as a hearing-impaired speaker.

8. ACKNOWLEDGEMENT

I am grateful to Esther Grabe-Georges for her help in the preparation of the manuscript.

9. REFERENCES

[1] ALTRINGHAM, J.D. & JOHNSTON, I.A. (1990) "Modelling muscle power output in a swimming fish", *J.Exp.Biol.,148*, 395-402.

[2] BADIN, P. (1989) "Acoustics of voiceless fricatives: production theory and data", *STL-QPSR 3/1989*, 33-55.

[3] COKER, C.H. (1976) "A model of articulatory dynamics and control", *Proc.IEEE, 64*, 452-460.

[4] GRACCO, V.L. (1988) "Timing factors in the coordination of speech movement", *J.of Neuro-science, 8*, 4628-4639.

[5] HARDCASTLE, W.J. & MARCHAL, A., Eds., (1990) *"Speech Production and Speech Modelling"*, Kluwer, Dordrecht.

[6] MILNER, T.E. (1986) "Controlling velocity in rapid movements", *J.of Motor Behaviour, 18*, 147-161.

[7] RAMOS, C.F. & STARK, L.W. (1987) "Simulation studies of descending and reflex control of fast movements", *J.of Motor Behaviour, 19*, 38-61.

[8] SCULLY, C. (1976) "A synthesizer study of aerodynamic factors in speech segment durations", in FANT, G., Ed. *"Progress in Speech Communication" Vol.2,* 227-234, Wiley, Stockholm.

[9] SCULLY, C. (1987) "Linguistic units and units of speech production", *Speech Comm., 6*, 77-142.

[10] SCULLY, C. & ALLWOOD, E. (1985) "Production and perception of an articulatory continuum for fricatives of English", *Speech Comm., 4*, 237-245.

[11] SHADLE, C.H. (1991) "The effect of geometry on source mechanisms of fricative consonants", *J.of Phonetics, in press.*

[12] STEVENS, K.N. (1971) "Airflow and turbulence noise for fricative and stop consonants: static considerations", *J.Acoust.Soc Amer., 50*, 1180-1192.